HAL
open science

A textometrical analysis of French arts workers "fr.Intermittents" on Twitter

Julien Longhi, Dalia Saigh

To cite this version:

Julien Longhi, Dalia Saigh. A textometrical analysis of French arts workers "fr.Intermittents" on Twitter. 4th conference CMC and Social Media Corpora for the Humanities, Sep 2016, Ljubljana, Slovenia. hal-01359511

HAL Id: hal-01359511

https://hal.science/hal-01359511v1

Submitted on 5 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Textometrical Analysis of French Arts Workers "fr.*Intermittents*" on Twitter

Julien Longhi, Dalia Saigh
Cergy-Pontoise University, AGORA
E-mail: julien.longhi@u-cergy.fr, dalia-saigh@hotmail.com

Abstract

The term "social media" is increasingly used and tends to replace the term Web 2.0. Through social networks, people create various relationships. The aim of this paper is to describe how communities of users interact with each other on a specific subject, especially on Twitter. The theme that we will study is about the controversy concerning French arts workers (*fr.intermittents*). We will conduct a textometrical analysis using the software Iramuteq and then explain the statistical results.

Keywords: social media, Twitter, *intermittents*, textometrical analysis, Iramuteq

1. Introduction

The term "social media" is increasingly used and tends to replace the term Web 2.0. Through social networks, people interact and create various relationships. In their exchanges, they establish content, organize, modify, and combine it with personal creations. Despite authors' freedom of expression and drafting, the content structure must obey rules of writing that are specific to each medium.

The aim of this paper is to analyze and describe how communities of users interact with each other on a specific subject. In our study, the theme is the controversy concerning French arts workers on Twitter: a microblogging service that is a hugely successful in spite of its particular working principle: blogging through ultra-short messages containing 140 characters. This feature allows the information flow faster but requires authors to be very concise when writing the tweet.

We will first describe the context and methodology for building our corpus. Then, we will introduce the method that we adopted for the textual analysis of this corpus entitled #intermittent (arts workers). We will also present Iramuteq, an analytical software tool that we have selected for this purpose and explain certain statistical results achieved.

2. Corpus Building: Background and Methodology

In March 2014, social partners signed a new agreement concerning the unemployment benefits for French arts workers. This text that became the convention of 14 May 2014 on unemployment benefits aroused concerns and opposition among the arts workers. A protest movement and mass demonstrations took place in Paris and in other French cities and lasted for several days.

These reactions rapidly invaded social networks especially Twitter. Millions of tweets were written as soon as the first information about this controversy emerged.

2.1 The Project Goal

The finalization of this corpus was made possible thanks to financial support from Ortolang[1]. The funding request centred around the finalization of the corpus-building process. The corpus is composed of tweets formed from the word hashtag (#) followed by the word "arts workers" then listed in a database of 13 074 tweets with #*intermittent*(s) and distributed in 4 617 twittos (Twitter users) over the period of June to September 2014, when tensions stepped up a notch and movements intensified.

Through the constitution of the corpus #*intermittent*, we hope to obtain a corpus which enables us to work on this kind of discourse (tweets related to a controversial topic), to characterize it and understand it under different forms in order to extend previous research (Longhi 2006, 2008) that focused on French arts workers in 2003/2004.

2.2 Data Building: the Choice of Data

After having contacted Twitter and having obtained confirmation that we had the right to collect and use information available on the site[2], we started tweet collection. This step was guided by the following process:

In 2014: retrieval of 13 074 tweets with #*intermittent* posted by 4 617 people.

In 2015: we established a threshold of at least 10 tweets with #*intermittent*: we obtained 215 accounts that had produced at least 10 tweets explicitly referenced as belonging to this theme (in order to have representative accounts). By collecting all the tweets from these 215 people, we gathered 586 239 tweets that included 10 876 tweets with #*intermittent*. The corpus #*intermittent* corresponds to these 10 876 tweets.

For the proper conduct of this process, we made, in collaboration with project participants from the field of Computing (Boris Borzic and AbdulhafizAlkhouli) a selection of data and metadata. For this, our colleagues developed a customized application. The application

1) uses the Twitter API: using ten functions of the API according to our needs, and recovering all the information in JSON format that we then convert;

2) allows the database to be enriched with a clean basic design (ten tables, fifty fields). Then we have programs that calculate indices for enriching additional fields;

3) allows customized export, with the information stored in a range of data formats. The challenge for a linguistic approach is to use this material to develop the #intermittent corpus.

These tweets were then formatted in TEI (with CMC formats extension tracks offered by a European group) to

[1] https://repository.ortolang.fr

[2] http://scinfolex.com/2009/06/14/twitter-et-le-droit-dauteur-vers-un-copyright-2-0/

become a corpus in order to meet the institutional requirements of the CoMeRe[3] project, and allow us to carry out a discourse analysis with word-processing tools on the corpus *#intermittent* or future corpora.

3. Textometrical Analysis of the Corpus *#Intermittent*

Textometry offers an instrumented approach to corpus analysis, articulating quantitative syntheses and analyzes including text (Lebart & Salem, 1994). Functionally, textometry implements differential principles. The approach highlights similarities and differences observed in the corpus according to the representation dimensions considered (lexical, grammatical, phonetic, or prosodic ones, etc). In addition to provide sorting procedures and statistical calculations for the study of digital corpora of texts, textometry establishes contextual and contrastive modeling. Thus, the text is characterized by its words in relation to their use in the corpus, the word is characterized by its co-occurrences, etc. (Pincemin, 2011).

Textometry is particularly relevant to corpus exploitation in human and social sciences. It simultaneously enables a detailed and global observation of different texts while remaining close to them, and highlights the fact that language is an important observation field for human and social sciences.

3.1 Iramuteq: the Text Analysis Tool

The Iramuteq software offers a set of analysis procedures for the description of a textual corpus. One of its principal methods is Alceste. This allows a user to segment a corpus into “context units”, to make comparisons and groupings of the segmented corpus according to the lexemes contained within it, and then to seek “stable distributions” (Reinert, 1998). In addition to the Alceste method, Iramuteq provides other analysis tools including prototypical analysis, similarities analysis, and word clouds analysis. All of these methods allow the users of this tool to map out the dynamics of the discourses of the different subjects engaged in interaction (Reinert, 1999).

3.2 The Corpus Structure

Input files for *Iramuteq* must be in text format (.txt) and observe the following formatting rules:

The basic unit is called "text". A text can represent an interview, an article, a book or any other type of documents. A corpus may contain one or more texts (but at least one). The texts are introduced by four stars (****) followed by a series of starred variables separated by a space. It is possible to put the starred variables within the text by introducing the beginning of the line by a hyphen followed by a star (- *). This is known as "themes". The line should contain only this variable.

For our corpus format, we have chosen a format with three representative variables: we called the first "*intermittent*", because it constitutes the key word of this corpus. The second is about the "usernames", it’s why this variable will change from a tweet to another depending on who posted the tweet. The third variable allows "the number" of tweets sent by a twittos to be counted as well as the re-tweets.

The figure below shows the formatting of the corpus *#intermittent*:

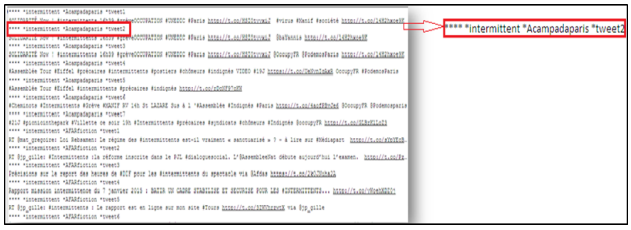

Figure 1: The format of the corpus *#intermittent*.

4. Methods and Results of the Analysis

4.1 The Word-Cloud

Iramuteq contains an option that makes a kind of a lexical compendium of a document in which the discussed key concepts are represented by a size unit (in the sense of the used typography weight). This allows their importance within the corpus to be highlighted. Specifically, the more a keyword is quoted in an article, the bigger it will appear in the cloud of words. This technique will allow us to put forward the keywords used by twittos.

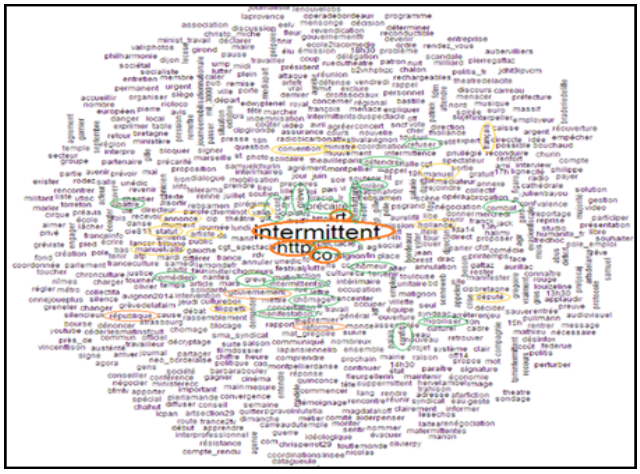

Figure 2: The word-cloud of the corpus *#intermittent*.

This word-cloud highlights the most common occurrences in tweets. These lexical items are positioned centrally in the cloud. The occurrence "*intermittent*" is the largest in size because it constitutes the key word of our corpus; this is why its frequency is higher. That word is followed by specific markers such as "co" and "http" that refer to links shared on Twitter. Indeed, these links are automatically abbreviated http: // co to allow long URLs to be shared without exceeding the maximum number of characters allowed when writing a tweet. There is also the sign "rt" which means "retweet". This has the function of reposting the tweet of another person enabling users to quickly share it with all subscribers.

[3]http://corpuscomere.wordpress.com

Around those keywords are others which have more or less the same frequency and thus appear the same size. Among them, those that refer to the semantic field of the republic and the French government such as *Manuel Valls*, *Republique* (republic), *député* (deputy), *français* (French), *F.Hollande*, *Fillipetit, minister* (minister). Other lexical items evoke either movements or activities such as *accord* (agreement), *grève* (strike), *mobilisation* (mobilization), *manifestation* (protest), *convention* (convention), *combat* (fight). There are also names or adjectives referring to French arts workers, and describing their situation as *chômeurs* (unemployed workers), *précaires* (precarious), *interluttents*, *comédiens* (actors).
Despite the interest of this method, the resulting description remains very general. For a more detailed analysis, *Iramuteq* offers another graphical representation of a corpus' words, a significant method called "similarities analysis", which retains the idea of size proportional to the frequency, but introduces the relations of co-occurrences between words.

4.2 Similarities Analysis

Similarities analysis is a technique based on graph theory (Flament, 1962). It presents in a graphical format the structure of a corpus, distinguishing between the shared parts and the specificities of coded variables. This allows the link between the different forms in the text segments to emerge (Marchand & Ratinaud, 2012).

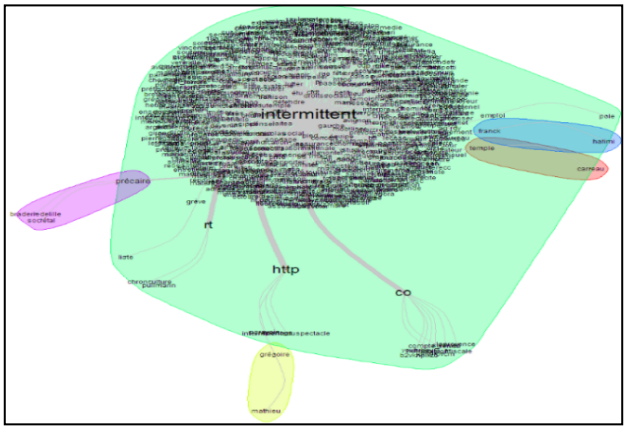

Figure 3: The similarities analysis of the corpus *#intermittent*.

The first observation that we can make is that this corpus is very homogeneous with one central idea around which revolves the greatest part of the lexicon of our corpus. This figure shows a single main cluster, with some others which are very small and not relevant. This cluster consists of a word cloud which contains the key word "*intermittent*" at its center and around it, are grouped a very dense and related lexicon.
We notice the presence of some small groups, which are in the main cluster, directly related (with edges) to "*intermittent*", the most important one. Among these groups, there is: "*http*" in which we find the term *intermittentdespectacle* (arts workers) and a little further, a small cluster containing the name *Gregory Mathieu*, a sociologist who wrote a book with the title "*Les intermittents du spectacle. Enjeux d'un siècle de luttes*". So, in this group, we understand that the majority of links mentioned in tweets refer users to web pages where the name of the sociologist is mentioned.
There is also the "*rt*" group which includes the following terms: *chronculture, pullmarin, dinamopress, angelin*... which refer to the names of people who have retweeted the most. The "co" group is, as explained above, the abbreviated form of links on Twitter.
We can already understand from this figure that the *#intermittent* corpus contains a lot of links, retweets related to French arts workers, and it describes their various actions and their status (highlighted by the cluster *précaires* (precarious).
That being said, as the lexicon related to the keyword "*intermittent*" is very dense, the function similarities analysis has simply helped us to describe the nature and the main topic of tweets (tweets with links, retweets, arts workers status ...). To further clarify the corpus structure, we will use the HDC "Hierarchical Descending Classification" function (a method established by Max Reinert).

4.3 The Hierarchical Descending Classification

One method used by Alceste is the hierarchical descending classification. This method offers a global approach to a corpus. The *HDC* after partitioning the corpus, identifies statistically independent word classes (forms). These classes are interpreted through their profiles, which are characterized by specific correlated forms. The *HDC* shows that using a dendrogram.

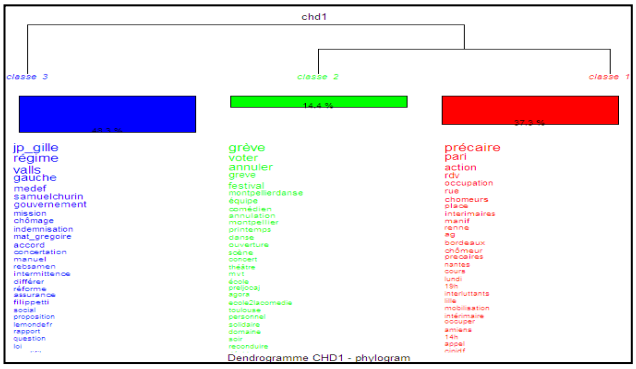

Figure 4: The result of the Hierarchical Descending Classification.

Two groups are distinguished in this figure, the first with two related classes (class 1 and class 2), and the second where there is only one class (class 3).
The class 1 includes forms associated with the different protest movements of French arts workers such as the occupation of streets, theaters and other places, the demonstrations in Paris and elsewhere.

Here is an extract of characteristic segments (with a high score), which contain the most common words associated with class 1 like *manif* (event), *cipdfjournée* (cip-idf day), *action* (action), the common words are highlighted in red:

**** *intermittent *CQFjournal *tweet10
score :1458.88
rt cipidf journée d action paris 10h république 14h manif ministère-du travail 127 rue de grenelle intermittents précaires htt t

**** *intermittent *CIP_IDF *tweet782
score : 1431.31
Rvd paris journée d actions coordonées 11h devant bourse du travail 3 rue du château d eau intermittents précaires httt co oz3kijtjuc

Figure 5: The characteristics segments of class 1.

Class 2 refers to strikes held by the French arts workers and their different concerts and show cancellations. This class contains words such as: *grève* (strike), *festival* (festival), *annulé* (cancelled). The following figure shows the characteristics segments of this class:

**** *intermittent *CIP_LR* tweet155
score :2058.76
intermittents rencontres photos arles la grève a été votée pour lundi 7 juillet jour de l ouverture du festival le vernissage annulé

**** *intermittent *cie813* tweet48
score :1877.02
second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

Figure 6: The characteristics segments of class 2.

Class 3 concerns the tweets that talk about the unemployment insurance system related to the French arts workers and political entities involved in this affair. Here is a characteristics segment summarizing the words associated with this class, including *medef*, *valls*, *samuelchurin*, *aurelifil*:

**** *intermittent *cie813* tweet48
score :1877.02
second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

**** *intermittent *AFARfiction *tweet42
score :403.83
rt jp_gille intermittents je viens de remettre mon rapport à manuel-valls premier ministre avec aurelifil et frebsamen httt cocb

Figure 7: The characteristics segments of class 3.

These results demonstrate that unlike the written press which showed a plurality of views concerning the semantic representation of the word "*intermittent*" (see Longhi, 2006) which was seen whether as a status (*statut*), a profession (*métier*) or in the dynamics of these two semantic components. Here, the word "*intermittent*" is presented using three different senses "system" (*régime*), "status" (*statut*) and "fight" (*lutte*). This indicates that Twitter focuses on the status side and declines it by introducing the French arts workers insurance system (one way of looking at the status) or the consequence of this status (fight).

5. Conclusion

A Textometrical analysis of this corpus has allowed us to see how twittos have reacted to the announcement of the new unemployment insurance system related to French arts workers. Through the analysis of similarities, we have found that there were a lot of links pointing to this topic with references to the sociologist Mathieu Grégoire and his various texts, and also newspaper names and publications including *Le Monde*. There were also various retweets and thanks to this, the issue has become in a short time a "trending topic" on Twitter. This is due to the various markers such as #, URLs, the @ sign ... The Reneirt method (HDC) taught us that discourse around this subject is divided into two different sets. On the one hand, tweets that describe the precariousness of French arts workers and their various protest movements against the new regime. On the other hand, tweets denouncing the impartiality of the agreement, with links providing information about that act and citing various political personalities who were involved in the controversy.

6. References

Flament, C. (1962). L'analyse de similitude. *Cahiers du centre de recherche opérationnelle*, 4, pp. 63--97

Lebart, L., Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Longhi, J. (2006). De intermittent du spectacle à intermittent: de la représentation à la nomination d'un objet du discours. *Corela*, 4 (2). URL : http://corela.revues.org/457.

Longhi, J. (2008). Sens communs et dynamiques sémantiques : l'objet discursif intermittent. *Langages*, 170, pp. 109--124.

Marchand, P., Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles. JADT, 2012*, pp. 687--699.

Pincemin, B. (2011). Sémantique interprétative et textométrie. *Corpus*, 10. URL: http://corpus.revues.org/2121.

Reinert, M. (1998). Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. *Actes des 4èmes journées Internationales d'Analyse Statistiques des Données textuelles*. URL : http://lexicometrica.univ-paris3.fr/jadt/jadt1998/reinert.htm.

Reinert, M. (1999). Quelques interrogations à propos de l'objet d'une analyse de discours de type statistique et de la réponse « Alceste ». *Langage et société*, 90 (1), pp. 57--70.

Corpus CoMeRe: https://corpuscomere.wordpress.com

Iramuteq: www.iramuteq.org

Ortolang: https://www.ortolang.fr/market/home